Milestone Completion for the LFSCK 2 Subproject 3.1.5 on the Lustre* software FSCK Project of the SFS-DEV-001 contract.

Revision History

Date	Revision	Author
2013-12-11	Original	R. Henwood

Contents

Introduction

The following milestone completion document applies to Subproject 3.2 – LFSCK 2: MDT-OST Consistency of the Lustre* FSCK within Amendment No. 1 on the OpenSFS Lustre Development contract SFS-DEV-001 agreed October 10, 2012.

Subproject Description

Per the contract, Implementation milestone is described as follows:

> *MDT-OST consistency will implement functionality for distributed verification and repair of the MDT inode to OST object mapping. This will add additional functionality while the MDT is iterating over the inodes (see Subproject 3.1) to check the file layout (LOV EA) to verify that the objects referenced in the file layout exist and that each object has a back reference to the correct MDT inode. Incorrect or missing back pointers on the OST objects will be corrected, and missing objects will be recreated when detected.*
>
> *The UID and GID of OST objects will also be verified to match that of the MDT inode to ensure correct quota allocation. After the MDT iteration is complete, any unreferenced OST objects will be linked into a lost+found directory.*
>
> *Subprojects 3.1 and 3.2 together constitute a complete replacement of the existing LFSCK utility for local file systems. This will allow complete checking of non-DNE file systems while the file system is online.*

Milestone Completion Criteria

Per the contract, Implementation milestone is described as follows:

> *Contractor shall complete implementation and unit testing for the approved solution. Contractor shall regularly report feature development progress including progress metrics at project meetings and engineers shall share interim unit testing results as they are available. OpenSFS at its discretion may request a code review. Completion of the implementation phase shall occur when the agreed to solution has been completed up to and including unit testing and this functionality can be demonstrated on a test cluster. Code Reviews shall include:*

---

* Other names and brands maybe the property of others.

a. *Discussion led by Contractor engineer providing an overview of Lustre source code changes*
b. *Review of any new unit test cases that were developed to test changes*

Location of Completed Solution

The agreed solution has been completed and is recorded in the following patches:

ID	Description
2a271b4	LU-3336 lfsck: create new MDT-object or exchange OST-objects
ac78f55	LU-3336 lfsck: recreate the lost MDT-object
416839c	LU-3336 lfsck: namespace visible lost+found directory
c61c112	LU-3336 lfsck: regenerate lost layout EA
56fdc3c	LU-3336 lfsck: orphan OST-objects iteration (2)
5b4f73b	LU-3336 lfsck: orphan OST-objects iteration (1)
4425fa1	LU-3336 lfsck: use rbtree to record OST-object accessing
d628a95	LU-3951 lfsck: OST-object inconsistency self detect/repair
aa87bcd	LU-3592 lfsck: repair multiple referenced OST-object
80054e6	LU-3594 lfsck: repair inconsistent OST-object owner
74c1059	LU-3591 lfsck: repair unmatched MDT-OST objects pairs
aa1e2e7	LU-3590 lfsck: repair MDT-object with dangling reference
cd21da7	LU-3593 lfsck: repair inconsistent layout EA
cc581f3	LU-1267 lfsck: enhance API for MDT-OST consistency
bf15de9	LU-3950 lfsck: control LFSCK on all devices via single command
7a998f3	LU-1267 lfsck: enhance RPCs (2) for MDT-OST consistency
a956e98	LU-1267 lfsck: enhance RPCs (1) for MDT-OST consistency
edb1bd9	LU-1267 lfsck: framework (3) for MDT-OST consistency
b3e6eda	LU-3951 lfsck: LWP connection from OST-x to MDT-y
f946d82	LU-1267 lfsck: framework (2) for MDT-OST consistency
141a375	LU-1267 lfsck: rebuild LAST_ID
ff36f47	LU-1267 lfsck: framework (1) for MDT-OST consistency
3e09d63	LU-3335 osd: use local transaction directly inside OSD
591f15e	LU-4106 scrub: Trigger OI scrub properly
f792da4	LU-3569 ofd: packing ost_idx in all IDIF

New functional tests

New functional tests to automatically verify the acceptance criteria agreed in the LFSCK2 Solution Architecture are available. The tests are contained in modifications to the files:

lustre/tests/sanity-lfsck.sh
lustre/tests/test-framework.sh

Demonstration of LFSCK 2 functionality.

Functional testing was completed on 2013-09-23. The detailed results are recorded in Appendix A. Section 5 of the LFSCK 2 Solution Architecture describes the acceptance tests including:

Acceptance test	**Corresponding code test from Appendix A**
5.1 Start/stop MDT-OST consistency check/repair through userspace commands	`test 0: Control LFSCK manually`
5.2 Monitor MDT-OST consistency check/repair	Present in all tests
5.3 Resume MDT-OST consistency check/repair from the latest checkpoint	`test 6a: LFSCK resumes from last checkpoint (1)` `test 6b: LFSCK resumes from last checkpoint (2)`
5.4 Rate control for MDT-OST consistency check/repair	`test 9a: LFSCK speed control (1)` `test 9b: LFSCK speed control (2)`
5.5 Repair file which parent has dangling reference	`test 13: LFSCK can repair MDT-object with dangling reference`
5.6 Repair unreferenced OST-object	`test 18a: OST-object inconsistency self detect` `test 18b: OST-object inconsistency self repair`
5.7 Repair unmatched referenced MDT-object and OST-object pair	`test 14a: LFSCK can repair unmatched MDT-object/OST-object pair (1)` `test 14b: LFSCK can repair unmatched MDT-object/OST-object pair (2)`
5.8 Repair repeated referenced OST-object	`test 16: LFSCK can repair multiple references`
5.9 Repair inconsistent file owner information	`test 15a: LFSCK can repair inconsistent MDT-object/OST-object pairs (1)`

	`test 15b: LFSCK can repair inconsistent MDT-object/OST-object pairs (2)`
5.10 Handle the OST upgrading from Lustre 1.8	`Beyond scope of test suite`
5.11 The Lustre system is available during the LFSCK for MDT-OST consistency.	`test 10: System is available during LFSCK scanning`

Conclusion

Implementation has been completed according to the agreed criteria.

Appendix A: functional test results from 2013-09-30

```
Logging to shared log directory: /tmp/test_logs/1382298292
RHEL6: Checking config lustre mounted on /mnt/lustre
Checking servers environments
Checking clients RHEL6 environments
Using TIMEOUT=20
disable quota as required
osd-ldiskfs.track_declares_assert=1
excepting tests:

== sanity-lfsck test 0: Control LFSCK manually == 03:44:53 (1382298293)
formatall
setupall
preparing... 10 * 10 files will be created.
prepared.
stop mds1
start mds1
fail_val=3
fail_loc=0x1600
Started LFSCK on the device lustre-MDT0000: namespace.
name: lfsck_namespace
magic: 0xa0629d03
version: 2
status: scanning-phase1
flags:
param:
time_since_last_completed: N/A
time_since_latest_start: 0 seconds
time_since_last_checkpoint: N/A
latest_start_position: 13, N/A, N/A
last_checkpoint_position: N/A, N/A, N/A
first_failure_position: N/A, N/A, N/A
checked_phase1: 0
checked_phase2: 0
updated_phase1: 0
updated_phase2: 0
failed_phase1: 0
failed_phase2: 0
dirs: 0
M-linked: 0
nlinks_repaired: 0
lost_found: 0
```

```
success_count: 0
run_time_phase1: 0 seconds
run_time_phase2: 0 seconds
average_speed_phase1: 0 items/sec
average_speed_phase2: N/A
real-time_speed_phase1: 0 items/sec
real-time_speed_phase2: N/A
current_position: 12, N/A, N/A
Stopped LFSCK on the device lustre-MDT0000.
Started LFSCK on the device lustre-MDT0000: namespace.
fail_loc=0
fail_val=0
Started LFSCK on the device lustre-MDT0000: namespace.
stopall, should NOT crash LU-3649
Resetting fail_loc on all nodes...done.
PASS 0 (78s)

== sanity-lfsck test 1a: LFSCK can find out and repair crashed FID-in-dirent ==
03:46:11 (1382298371)
formatall
setupall
preparing... 1 * 1 files will be created.
prepared.
stop mds1
start mds1
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
fail_loc=0x1501
fail_loc=0
10.211.55.5@tcp:/lustre /mnt/lustre lustre rw,flock,user_xattr 0 0
Stopping client RHEL6 /mnt/lustre (opts:)
Started LFSCK on the device lustre-MDT0000: namespace.
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
fail_loc=0x1505
fail_loc=0
Resetting fail_loc on all nodes...done.
PASS 1a (45s)

== sanity-lfsck test 1b: LFSCK can find out and repair missed FID-in-LMA ==
03:46:56 (1382298416)
formatall
setupall
preparing... 1 * 1 files will be created.
prepared.
stop mds1
start mds1
```

```
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
fail_loc=0x1502
fail_loc=0
10.211.55.5@tcp:/lustre /mnt/lustre lustre rw,flock,user_xattr 0 0
Stopping client RHEL6 /mnt/lustre (opts:)
fail_loc=0x1506
Started LFSCK on the device lustre-MDT0000: namespace.
fail_loc=0
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
fail_loc=0x1505
fail_loc=0
Resetting fail_loc on all nodes...done.
PASS 1b (65s)

== sanity-lfsck test 2a: LFSCK can find out and repair crashed linkEA entry ==
03:48:01 (1382298481)
formatall
setupall
preparing... 1 * 1 files will be created.
prepared.
stop mds1
start mds1
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
fail_loc=0x1603
fail_loc=0
10.211.55.5@tcp:/lustre /mnt/lustre lustre rw,flock,user_xattr 0 0
Stopping client RHEL6 /mnt/lustre (opts:)
Started LFSCK on the device lustre-MDT0000: namespace.
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
Resetting fail_loc on all nodes...done.
PASS 2a (66s)

== sanity-lfsck test 2b: LFSCK can find out and remove invalid linkEA entry ==
03:49:07 (1382298547)
formatall
setupall
preparing... 1 * 1 files will be created.
prepared.
stop mds1
start mds1
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
fail_loc=0x1604
fail_loc=0
10.211.55.5@tcp:/lustre /mnt/lustre lustre rw,flock,user_xattr 0 0
Stopping client RHEL6 /mnt/lustre (opts:)
```

```
Started LFSCK on the device lustre-MDT0000: namespace.
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
Resetting fail_loc on all nodes...done.
PASS 2b (58s)

== sanity-lfsck test 2c: LFSCK can find out and remove repeated linkEA entry ==
03:50:05 (1382298605)
formatall
setupall
preparing... 1 * 1 files will be created.
prepared.
stop mds1
start mds1
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
fail_loc=0x1605
fail_loc=0
10.211.55.5@tcp:/lustre /mnt/lustre lustre rw,flock,user_xattr 0 0
Stopping client RHEL6 /mnt/lustre (opts:)
Started LFSCK on the device lustre-MDT0000: namespace.
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
Resetting fail_loc on all nodes...done.
PASS 2c (61s)

== sanity-lfsck test 4: FID-in-dirent can be rebuilt after MDT file-level
backup/restore == 03:51:06 (1382298666)
formatall
setupall
preparing... 3 * 3 files will be created.
prepared.
stop mds1
file-level backup/restore on mds1:/tmp/lustre-mdt1
backup EA
/root/Work/Lustre/L80/lustre-release/lustre/tests
backup data
reformat new device
restore data
restore EA
/root/Work/Lustre/L80/lustre-release/lustre/tests
remove recovery logs
removed `/mnt/brpt/CATALOGS'
start mds1 with disabling OI scrub
fail_val=1
fail_loc=0x1601
Started LFSCK on the device lustre-MDT0000: namespace.
fail_loc=0
```

```
fail_val=0
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
fail_loc=0x1505
fail_loc=0
Resetting fail_loc on all nodes...done.
PASS 4 (57s)

== sanity-lfsck test 5: LFSCK can handle IFIG object upgrading == 03:52:03
(1382298723)
formatall
setupall
fail_loc=0x1504
preparing... 1 * 1 files will be created.
fail_loc=0
prepared.
stop mds1
file-level backup/restore on mds1:/tmp/lustre-mdt1
backup EA
/root/Work/Lustre/L80/lustre-release/lustre/tests
backup data
reformat new device
restore data
restore EA
/root/Work/Lustre/L80/lustre-release/lustre/tests
remove recovery logs
removed `/mnt/brpt/CATALOGS'
start mds1 with disabling OI scrub
fail_val=1
fail_loc=0x1601
Started LFSCK on the device lustre-MDT0000: namespace.
fail_loc=0
fail_val=0
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
fail_loc=0x1505
fail_loc=0
Resetting fail_loc on all nodes...done.
PASS 5 (59s)

== sanity-lfsck test 6a: LFSCK resumes from last checkpoint (1) == 03:53:02
(1382298782)
formatall
setupall
preparing... 10 * 10 files will be created.
prepared.
stop mds1
```

```
start mds1
fail_val=1
fail_loc=0x1600
Started LFSCK on the device lustre-MDT0000: namespace.
fail_loc=0x80001608
fail_val=1
fail_loc=0x1600
Started LFSCK on the device lustre-MDT0000: namespace.
fail_loc=0
fail_val=0
Resetting fail_loc on all nodes...done.
PASS 6a (62s)

== sanity-lfsck test 6b: LFSCK resumes from last checkpoint (2) == 03:54:04
(1382298844)
formatall
setupall
preparing... 10 * 10 files will be created.
prepared.
stop mds1
start mds1
fail_val=1
fail_loc=0x1601
Started LFSCK on the device lustre-MDT0000: namespace.
fail_loc=0x80001609
fail_val=1
fail_loc=0x1601
Started LFSCK on the device lustre-MDT0000: namespace.
fail_loc=0
fail_val=0
Resetting fail_loc on all nodes...done.
PASS 6b (64s)

== sanity-lfsck test 7a: non-stopped LFSCK should auto restarts after MDS remount
(1) == 03:55:08 (1382298908)
formatall
setupall
preparing... 10 * 10 files will be created.
prepared.
stop mds1
start mds1
fail_val=1
fail_loc=0x1601
Started LFSCK on the device lustre-MDT0000: namespace.
stop mds1
```

```
start mds1
fail_loc=0
fail_val=0
Resetting fail_loc on all nodes...done.
PASS 7a (61s)

== sanity-lfsck test 7b: non-stopped LFSCK should auto restarts after MDS remount
(2) == 03:56:09 (1382298969)
formatall
setupall
preparing... 2 * 2 files will be created.
prepared.
stop mds1
start mds1
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
fail_loc=0x1604
fail_val=1
fail_loc=0x1602
Started LFSCK on the device lustre-MDT0000: namespace.
stop mds1
start mds1
fail_loc=0
fail_val=0
Resetting fail_loc on all nodes...done.
PASS 7b (67s)

== sanity-lfsck test 8: LFSCK state machine == 03:57:16 (1382299036)
formatall
setupall
preparing... 20 * 20 files will be created.
prepared.
stop mds1
start mds1
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
fail_loc=0x1603
fail_loc=0x1604
fail_val=2
fail_loc=0x1601
Started LFSCK on the device lustre-MDT0000: namespace.
Stopped LFSCK on the device lustre-MDT0000.
Started LFSCK on the device lustre-MDT0000: namespace.
fail_loc=0x80001609
fail_loc=0x1600
Started LFSCK on the device lustre-MDT0000: namespace.
fail_loc=0x160a
```

```
stop mds1
fail_loc=0x160b
start mds1
fail_loc=0x1601
Started LFSCK on the device lustre-MDT0000: namespace.
stop mds1
fail_loc=0x160b
start mds1
fail_val=2
fail_loc=0x1602
Started LFSCK on the device lustre-MDT0000: namespace.
fail_loc=0
fail_val=0
Resetting fail_loc on all nodes...done.
PASS 8 (76s)

== sanity-lfsck test 9a: LFSCK speed control (1) == 03:58:32 (1382299112)
formatall
setupall
preparing... 70 * 70 files will be created.
prepared.
stop mds1
start mds1
Started LFSCK on the device lustre-MDT0000: namespace.
Resetting fail_loc on all nodes...done.
PASS 9a (88s)

== sanity-lfsck test 9b: LFSCK speed control (2) == 04:00:00 (1382299200)
formatall
setupall
preparing... 0 * 0 files will be created.
prepared.
stop mds1
start mds1
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
Another preparing... 50 * 50 files (with error) will be created.
fail_loc=0x1604
fail_loc=0x160c
Started LFSCK on the device lustre-MDT0000: namespace.
fail_loc=0
Started LFSCK on the device lustre-MDT0000: namespace.
Resetting fail_loc on all nodes...done.
PASS 9b (94s)
```

```
== sanity-lfsck test 10: System is available during LFSCK scanning == 04:01:34
(1382299294)
formatall
setupall
preparing... 1 * 1 files will be created.
prepared.
stop mds1
start mds1
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
fail_loc=0x1603
fail_loc=0x1604
fail_loc=0
10.211.55.5@tcp:/lustre /mnt/lustre lustre rw,flock,user_xattr 0 0
Stopping client RHEL6 /mnt/lustre (opts:)
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
Started LFSCK on the device lustre-MDT0000: namespace.
10.211.55.5@tcp:/lustre /mnt/lustre lustre rw,flock,user_xattr 0 0
Stopping client RHEL6 /mnt/lustre (opts:)
Resetting fail_loc on all nodes...done.
PASS 10 (102s)

== sanity-lfsck test 11a: LFSCK can rebuild lost last_id == 04:03:16 (1382299396)
stopall
formatall
setupall
total: 64 creates in 0.14 seconds: 457.69 creates/second
stop ost1
remove LAST_ID: idx=0
removed `/mnt/brpt/O/0/LAST_ID'
start ost1
fail_val=3
fail_loc=0x160e
trigger LFSCK for layout on ost1 to rebuild the LAST_ID(s)
Started LFSCK on the device lustre-OST0000: layout.
fail_val=0
fail_loc=0
the LAST_ID(s) should have been rebuilt
Resetting fail_loc on all nodes...done.
PASS 11a (43s)

== sanity-lfsck test 11b: LFSCK can rebuild crashed last_id == 04:03:59
(1382299439)
stopall
formatall
setupall
```

```
set fail_loc=0x160d to skip the updating LAST_ID on-disk
fail_loc=0x160d
total: 64 creates in 0.14 seconds: 448.34 creates/second
10.211.55.5@tcp:/lustre /mnt/lustre lustre rw,flock,user_xattr 0 0
Stopping client RHEL6 /mnt/lustre (opts:)
stop ost1
Stopping /mnt/ost1 (opts:) on RHEL6
fail_val=0
fail_loc=0x215
start ost1
Starting ost1:   -o loop /tmp/lustre-ost1 /mnt/ost1
Started lustre-OST0000
the on-disk LAST_ID should be smaller than the expected one
trigger LFSCK for layout on ost1 to rebuild the on-disk LAST_ID
Started LFSCK on the device lustre-OST0000: layout.
stop ost1
Stopping /mnt/ost1 (opts:) on RHEL6
start ost1
Starting ost1:   -o loop /tmp/lustre-ost1 /mnt/ost1
Started lustre-OST0000
the on-disk LAST_ID should have been rebuilt
fail_loc=0
Resetting fail_loc on all nodes...done.
PASS 11b (63s)

== sanity-lfsck test 12: single command to trigger LFSCK on all devices == 04:05:02
(1382299502)
stopall
formatall
setupall
All the LFSCK targets should be in 'init' status.
total: 100 creates in 0.21 seconds: 465.90 creates/second
total: 100 creates in 0.21 seconds: 477.78 creates/second
Trigger LFSCK on all targets by single command (limited speed).
Started LFSCK on the device lustre-MDT0000: layout.
All the LFSCK targets should be in 'scanning-phase1' status.
Stop layout LFSCK on all targets by single lctl command.
Stopped LFSCK on the device lustre-MDT0000.
All the LFSCK targets should be in 'stopped' status.
Re-trigger LFSCK on all targets by single command (full speed).
Started LFSCK on the device lustre-MDT0000: layout.
All the LFSCK targets should be in 'completed' status.
Resetting fail_loc on all nodes...done.
PASS 12 (44s)
```

```
== sanity-lfsck test 13: LFSCK can repair crashed lmm_oi == 04:05:46 (1382299546)
#####
The lmm_oi in layout EA should be consistent with the MDT-object
FID; otherwise, the LFSCK should re-generate the lmm_oi from the
MDT-object FID.
#####
stopall
formatall
setupall
Inject failure stub to simulate bad lmm_oi
fail_loc=0x160f
total: 32 creates in 0.07 seconds: 454.86 creates/second
fail_loc=0
stopall to cleanup object cache
setupall
Trigger layout LFSCK to find out the bad lmm_oi and fix them
Started LFSCK on the device lustre-MDT0000: layout.
Resetting fail_loc on all nodes...done.
PASS 13 (86s)

== sanity-lfsck test 14: LFSCK can repair MDT-object with dangling reference ==
04:07:12 (1382299632)
#####
The OST-object referenced by the MDT-object should be there;
otherwise, the LFSCK should re-create the missed OST-object.
#####
stopall
formatall
setupall
Inject failure stub to simulate dangling referenced MDT-object
fail_loc=0x1610
total: 64 creates in 0.14 seconds: 443.73 creates/second
fail_loc=0
stopall to cleanup object cache
setupall
'ls' should fail because of dangling referenced MDT-object
Trigger layout LFSCK to find out dangling reference and fix them
Started LFSCK on the device lustre-MDT0000: layout.
'ls' should success after layout LFSCK repairing
Resetting fail_loc on all nodes...done.
PASS 14 (85s)

== sanity-lfsck test 15a: LFSCK can repair unmatched MDT-object/OST-object pairs
(1) == 04:08:37 (1382299717)
#####
```

```
If the OST-object referenced by the MDT-object back points
to some non-exist MDT-object, then the LFSCK should repair
the OST-object to back point to the right MDT-object.
#####
stopall
formatall
setupall
Inject failure stub to make the OST-object to back point to
non-exist MDT-object.
fail_loc=0x1611
1+0 records in
1+0 records out
1048576 bytes (1.0 MB) copied, 0.00272507 s, 385 MB/s
fail_loc=0
stopall to cleanup object cache
setupall
Trigger layout LFSCK to find out unmatched pairs and fix them
Started LFSCK on the device lustre-MDT0000: layout.
Resetting fail_loc on all nodes...done.
PASS 15a (88s)

== sanity-lfsck test 15b: LFSCK can repair unmatched MDT-object/OST-object pairs
(2) == 04:10:05 (1382299805)
#####
If the OST-object referenced by the MDT-object back points
to other MDT-object that doesn't recognize the OST-object,
then the LFSCK should repair it to back point to the right
MDT-object (the first one).
#####
stopall
formatall
setupall
Inject failure stub to make the OST-object to back point to
other MDT-object
fail_loc=0x1612
1+0 records in
1+0 records out
1048576 bytes (1.0 MB) copied, 0.00263859 s, 397 MB/s
fail_loc=0
stopall to cleanup object cache
setupall
Trigger layout LFSCK to find out unmatched pairs and fix them
Started LFSCK on the device lustre-MDT0000: layout.
Resetting fail_loc on all nodes...done.
PASS 15b (92s)
```

```
== sanity-lfsck test 16: LFSCK can repair inconsistent MDT-object/OST-object owner
== 04:11:37 (1382299897)
#####
If the OST-object's owner information does not match the owner
information stored in the MDT-object, then the LFSCK trust the
MDT-object and update the OST-object's owner information.
#####
stopall
formatall
setupall
Inject failure stub to skip OST-object owner changing
fail_loc=0x1613
total: 1 creates in 0.00 seconds: 733.14 creates/second
fail_loc=0
Trigger layout LFSCK to find out inconsistent OST-object owner
and fix them
Started LFSCK on the device lustre-MDT0000: layout.
Resetting fail_loc on all nodes...done.
PASS 16 (43s)

== sanity-lfsck test 17: LFSCK can repair multiple references == 04:12:20
(1382299940)
#####
If more than one MDT-objects reference the same OST-object,
and the OST-object only recognizes one MDT-object, then the
LFSCK should create new OST-objects for such non-recognized
MDT-objects.
#####
stopall
formatall
setupall
Inject failure stub to make two MDT-objects to refernce
the OST-object
fail_val=0
fail_loc=0x1614
1+0 records in
1+0 records out
1048576 bytes (1.0 MB) copied, 0.00299106 s, 351 MB/s
fail_loc=0
fail_val=0
stopall to cleanup object cache
setupall
/mnt/lustre/d0.sanity-lfsck/d17/f0 and /mnt/lustre/d0.sanity-lfsck/d17/guard use
the same OST-objects
```

```
Trigger layout LFSCK to find out multiple refenced MDT-objects
and fix them
Started LFSCK on the device lustre-MDT0000: layout.
/mnt/lustre/d0.sanity-lfsck/d17/f0 and /mnt/lustre/d0.sanity-lfsck/d17/guard should
use diff OST-objects
2+0 records in
2+0 records out
2097152 bytes (2.1 MB) copied, 0.00374863 s, 559 MB/s
Resetting fail_loc on all nodes...done.
PASS 17 (94s)

== sanity-lfsck test 18a: OST-object inconsistency self detect == 04:13:54
(1382300034)
stopall
formatall
setupall
10.211.55.5@tcp:/lustre /mnt/lustre lustre rw,flock,user_xattr 0 0
Stopping client RHEL6 /mnt/lustre (opts:)
Starting client: RHEL6: -o user_xattr,flock RHEL6@tcp:/lustre /mnt/lustre
Inject failure, then client will offer wrong parent FID when read
fail_loc=0x1615
Read RPC with wrong parent FID should be denied
cat: /mnt/lustre/d0.sanity-lfsck/d18/a0: Input/output error
fail_loc=0
Resetting fail_loc on all nodes...done.
PASS 18a (59s)

== sanity-lfsck test 18b: OST-object inconsistency self repair == 04:14:53
(1382300093)
stopall
formatall
setupall
Inject failure stub to make the OST-object to back point to
non-exist MDT-object
fail_loc=0x1611
fail_loc=0
Nothing should be fixed since self detect and repair is disabled
Read RPC with right parent FID should be accepted,
and cause parent FID on OST to be fixed
foo
Resetting fail_loc on all nodes...done.
PASS 18b (44s)

== sanity-lfsck test 19a: Find out orphan OST-object and repair it (1) == 04:15:37
(1382300137)
#####
```

```
The target MDT-object is there, but related stripe information
is lost or partly lost. The LFSCK should re-generate the those
missed layout EA entries.
#####
stopall
formatall
setupall
2+0 records in
2+0 records out
2097152 bytes (2.1 MB) copied, 0.00502541 s, 417 MB/s
2+0 records in
2+0 records out
2097152 bytes (2.1 MB) copied, 0.00504708 s, 416 MB/s
[0x280000400:0x6:0x0]
/mnt/lustre/d0.sanity-lfsck/d19/a1/f1
lmm_stripe_count:   1
lmm_stripe_size:    1048576
lmm_pattern:        1
lmm_layout_gen:     0
lmm_stripe_offset:  0
      obdidx             objid             objid             group
           0                 2                0x2                 0

[0x2c0000400:0x2:0x0]
/mnt/lustre/d0.sanity-lfsck/d19/a2/f2
lmm_stripe_count:   2
lmm_stripe_size:    1048576
lmm_pattern:        1
lmm_layout_gen:     0
lmm_stripe_offset:  1
      obdidx             objid             objid             group
           1                 2                0x2        0x240000400
           0                 2                0x2        0x200000400

Inject failure, to make the MDT-object lost its layout EA
fail_loc=0x1616
fail_loc=0
stopall to cleanup object cache
setupall
The file size should be incorrect since layout EA is lost
Trigger layout LFSCK on all devices to find out orphan OST-objects
Started LFSCK on the device lustre-MDT0000: layout.
[0x280000400:0x6:0x0]
/mnt/lustre/d0.sanity-lfsck/d19/a1/f1
lmm_stripe_count:   1
```

```
lmm_stripe_size:    1048576
lmm_pattern:        1
lmm_layout_gen:     0
lmm_stripe_offset:  0
      obdidx             objid             objid             group
           0                 2                0x2                 0

[0x2c0000400:0x2:0x0]
/mnt/lustre/d0.sanity-lfsck/d19/a2/f2
lmm_stripe_count:   2
lmm_stripe_size:    1048576
lmm_pattern:        1
lmm_layout_gen:     1
lmm_stripe_offset:  1
      obdidx             objid             objid               group
           1                 2                0x2        0x240000400
           0                 2                0x2        0x200000400

The file size should be correct after layout LFSCK scanning
Resetting fail_loc on all nodes...done.
PASS 19a (83s)

== sanity-lfsck test 19b: Find out orphan OST-object and repair it (2) == 04:17:00
(1382300220)
#####
The target MDT-object is lost. The LFSCK should re-create the
MDT-object under .lustre/lost+found/MDTxxxx. The admin should
can move it back to normal namespace manually.
#####
stopall
formatall
setupall
2+0 records in
2+0 records out
2097152 bytes (2.1 MB) copied, 0.00454672 s, 461 MB/s
2+0 records in
2+0 records out
2097152 bytes (2.1 MB) copied, 0.0050405 s, 416 MB/s
[0x280000400:0x6:0x0]
/mnt/lustre/d0.sanity-lfsck/d19/a1/f1
lmm_stripe_count:   1
lmm_stripe_size:    1048576
lmm_pattern:        1
lmm_layout_gen:     0
lmm_stripe_offset:  0
```

```
      obdidx             objid             objid             group
           0                 2                 0x2                 0

[0x2c0000400:0x2:0x0]
/mnt/lustre/d0.sanity-lfsck/d19/a2/f2
lmm_stripe_count:   2
lmm_stripe_size:    1048576
lmm_pattern:        1
lmm_layout_gen:     0
lmm_stripe_offset:  1
      obdidx             objid             objid             group
           1                 2                 0x2        0x240000400
           0                 2                 0x2        0x200000400

Inject failure, to simulate the case of missing the MDT-object
fail_loc=0x1617
fail_loc=0
stopall to cleanup object cache
setupall
Trigger layout LFSCK on all devices to find out orphan OST-objects
Started LFSCK on the device lustre-MDT0000: layout.
Move the files from ./lustre/lost+found/MDTxxxx to namespace
[0x280000400:0x6:0x0]
/mnt/lustre/d0.sanity-lfsck/d19/a1/f1
lmm_stripe_count:   1
lmm_stripe_size:    1048576
lmm_pattern:        1
lmm_layout_gen:     0
lmm_stripe_offset:  0
      obdidx             objid             objid             group
           0                 2                 0x2                 0

[0x2c0000400:0x2:0x0]
/mnt/lustre/d0.sanity-lfsck/d19/a2/f2
lmm_stripe_count:   2
lmm_stripe_size:    1048576
lmm_pattern:        1
lmm_layout_gen:     1
lmm_stripe_offset:  1
      obdidx             objid             objid             group
           1                 2                 0x2        0x240000400
           0                 2                 0x2        0x200000400

The file size should be correct after layout LFSCK scanning
Resetting fail_loc on all nodes...done.
```

```
PASS 19b (79s)

== sanity-lfsck test 19c: Find out orphan OST-object and repair it (3) == 04:18:19
(1382300299)
#####
The target MDT-object layout EA slot is occpuied by some new
created OST-object when repair dangling reference case. Then
the LFSCK will exchange the two conflict OST-objects in such
MDT-object layout EA.
#####
stopall
formatall
setupall
1+0 records in
1+0 records out
1048576 bytes (1.0 MB) copied, 0.00362708 s, 289 MB/s
1+0 records in
1+0 records out
1048576 bytes (1.0 MB) copied, 0.00259489 s, 404 MB/s
[0x280000400:0x5:0x0]
/mnt/lustre/d0.sanity-lfsck/d19/a1/f1
lmm_stripe_count:   1
lmm_stripe_size:    1048576
lmm_pattern:        1
lmm_layout_gen:     0
lmm_stripe_offset:  0
      obdidx             objid             objid             group
           0                 2                0x2                 0

[0x280000400:0x6:0x0]
/mnt/lustre/d0.sanity-lfsck/d19/a1/f2
lmm_stripe_count:   1
lmm_stripe_size:    1048576
lmm_pattern:        1
lmm_layout_gen:     0
lmm_stripe_offset:  0
      obdidx             objid             objid             group
           0                 3                0x3                 0

Inject failure, to make /mnt/lustre/d0.sanity-lfsck/d19/a1/f2 and
/mnt/lustre/d0.sanity-lfsck/d19/a1/f1
to reference the same OST-object. Then drop /mnt/lustre/d0.sanity-lfsck/d19/a1/f1.
So /mnt/lustre/d0.sanity-lfsck/d19/a1/f2 will dangling reference, although its old
OST-object still is there.
```

```
fail_loc=0x1618
fail_loc=0
stopall to cleanup object cache
setupall
The file size should be incorrect since dangling referenced
ls: cannot access /mnt/lustre/d0.sanity-lfsck/d19/a1/f2: Cannot allocate memory
Trigger layout LFSCK on all devices to find out orphan OST-objects
Started LFSCK on the device lustre-MDT0000: layout.
The file size should be correct after layout LFSCK scanning
[0x280000400:0x6:0x0]
/mnt/lustre/d0.sanity-lfsck/d19/a1/f2
lmm_stripe_count:   1
lmm_stripe_size:    1048576
lmm_pattern:        1
lmm_layout_gen:     1
lmm_stripe_offset:  0
      obdidx             objid             objid             group
           0                 3                0x3                 0

There should be some stub under .lustre/lost+found/MDT0000/
180144035829121026 -rwx------ 1 bin bin 0 Oct 21 04:19
/mnt/lustre/.lustre/lost+found/MDT0000/E-[0x280000bd0:0x2:0x0]-0
Resetting fail_loc on all nodes...done.
PASS 19c (82s)
Stopping clients: RHEL6 /mnt/lustre (opts:)
Stopping client RHEL6 /mnt/lustre opts:
Stopping clients: RHEL6 /mnt/lustre2 (opts:)
Stopping /mnt/mds1 (opts:-f) on RHEL6
Stopping /mnt/mds2 (opts:-f) on RHEL6
Stopping /mnt/ost1 (opts:-f) on RHEL6
Stopping /mnt/ost2 (opts:-f) on RHEL6
Loading modules from /root/Work/Lustre/L80/lustre-release/lustre
detected 2 online CPUs by sysfs
Force libcfs to create 2 CPU partitions
debug=vfstrace rpctrace dlmtrace neterror ha config ioctl super
subsystem_debug=all -lnet -lnd -pinger
gss/krb5 is not supported
Formatting mgs, mds, osts
Format mds1: /tmp/lustre-mdt1
Format mds2: /tmp/lustre-mdt2
Format ost1: /tmp/lustre-ost1
Format ost2: /tmp/lustre-ost2
== sanity-lfsck test complete, duration 2110 sec == 04:20:02 (1382300402)
```